NSF

The MSP-Conversation Corpus

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso

INTERSPEECH 2020

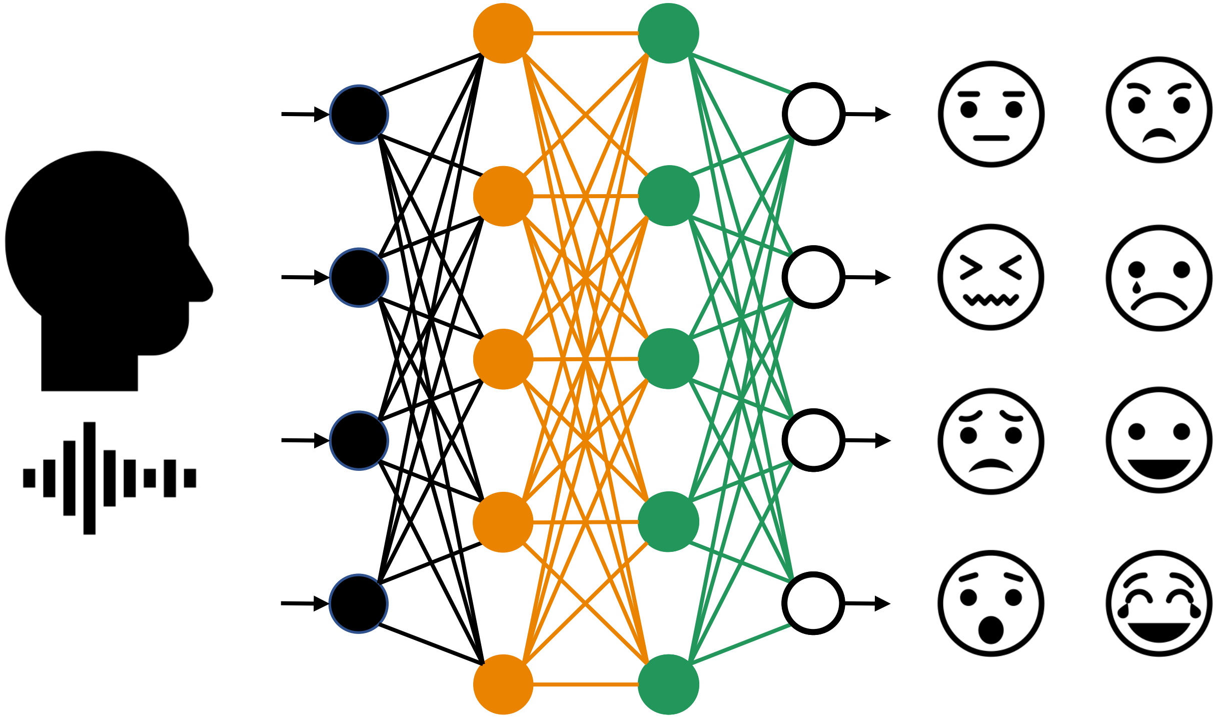
UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP Multimodal Signal Processing Laboratory

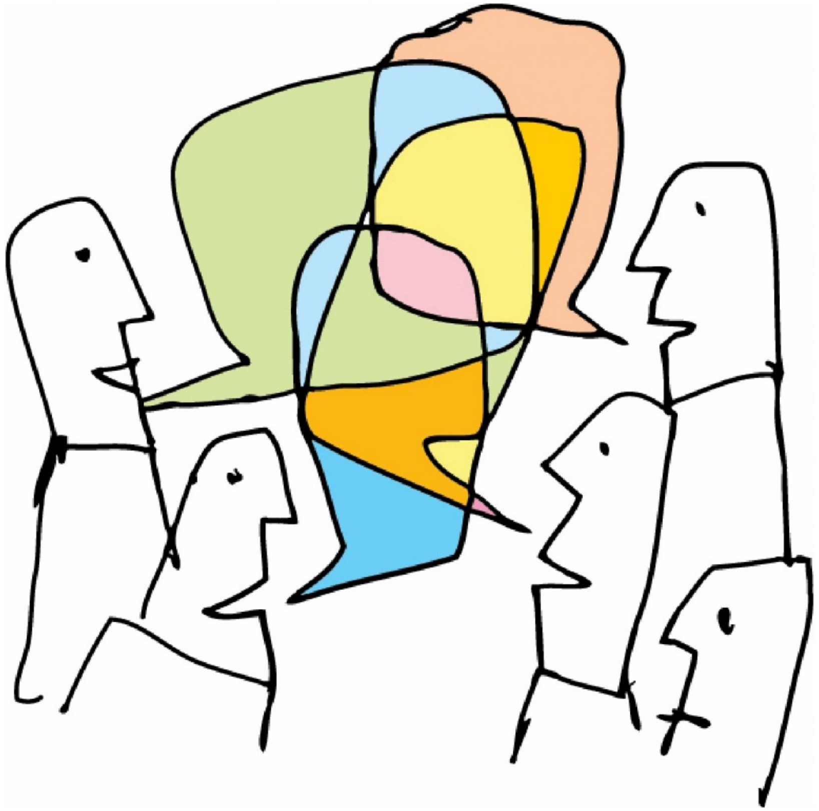
Introduction

- **Emotion recognition systems are needed for seamless human-computer interaction (HCI)**
  - Modality of speech is common
  - Effective speech emotion recognition (SER) systems
  - We need large amount of natural speech data annotated with emotional labels



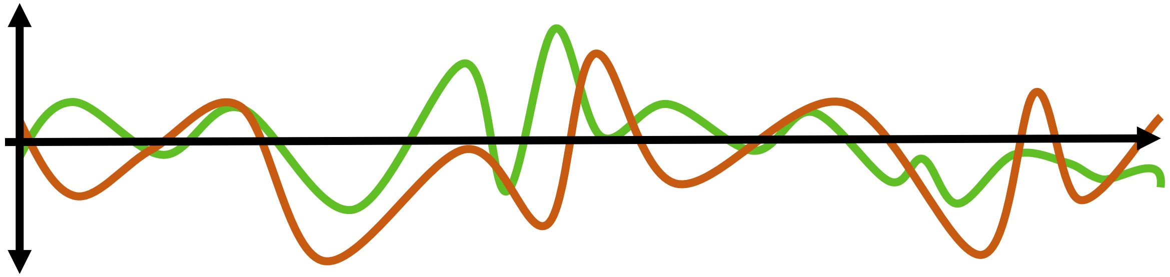
Introduction

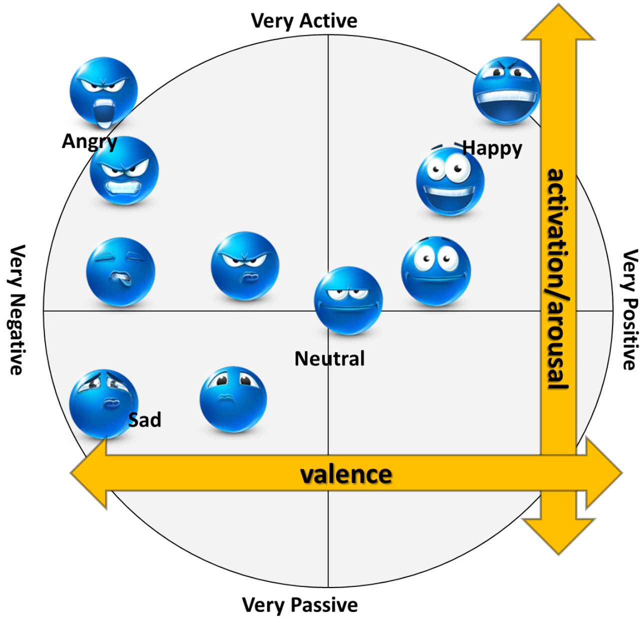
- **Most corpora are annotated without context at the sentence level**
  - Not appropriate for temporal modeling of emotion
- **We present the MSP-Conversation Corpus**
  - Naturalistic recordings obtained from online podcasts
    - Segments between 10 and 20 minutes long
  - Broad range of topics of conversations
  - Time-continuous annotations (emotional traces)



Introduction

- **Time-continuous annotations**
  - Emotions are dynamic and affected by contextual information
  - Instantaneous emotional perception of evaluators
  - We can study emotions at various temporal resolutions
- **Emotional Attributes**
  - Natural emotional behaviors are too complex for a finite number of classes
  - We use the emotional attributes:
    - Arousal (active versus calm)
    - Valence (positive versus negative)
    - Dominance (strong versus weak)





Outline of Presentation

1. Introduction
2. Related Work
3. The MSP-Conversation Corpus
4. Analysis of the Corpus
5. Conclusion
6. Future Work

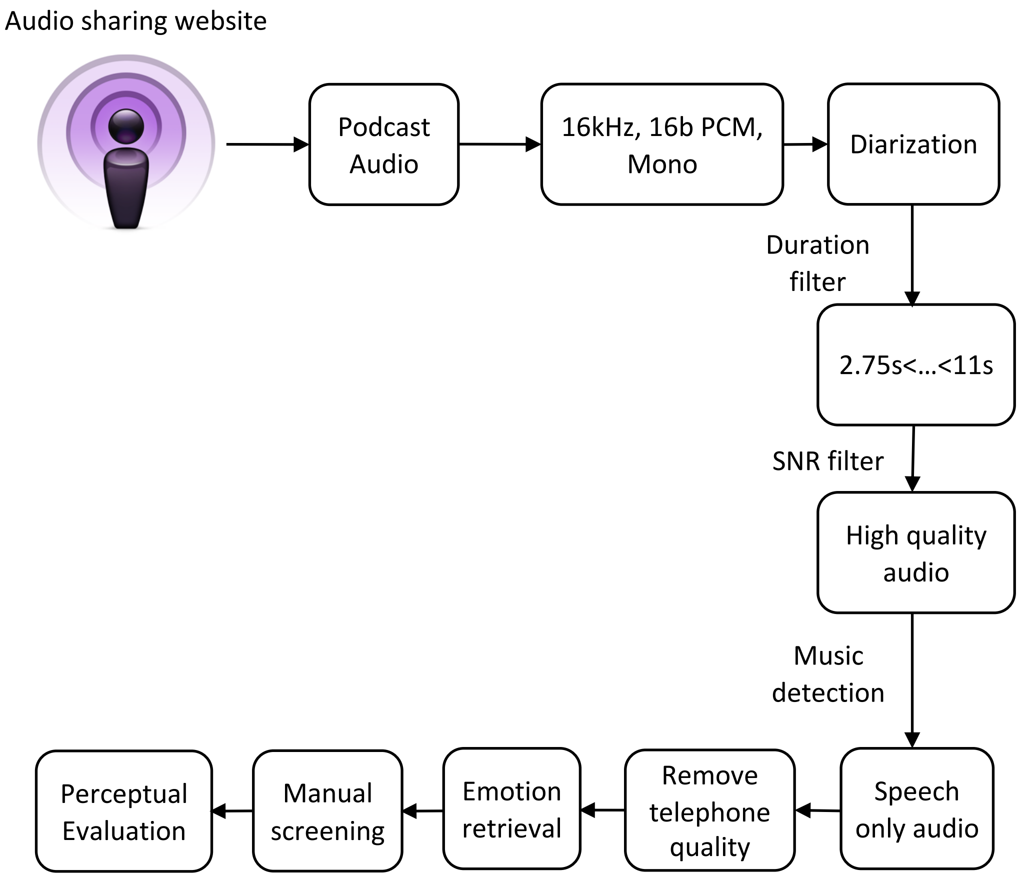
Related Work

- **Corpora with time-continuous annotations**
  - SEMAINE’s Solid SAL approach has an ‘operator’ who intends to induce emotions in the ‘user’
  - RECOLA and SEWA use emotional stimuli to induce emotional behaviors from participants
  - MuSe-Car uses in the wild recordings from one domain (car reviews)

Database	Type	Duration	Speakers
CreativeIT	Acted	≈8hrs	16
SEMAINE	Natural	15.83hrs	28
RECOLA	Natural	3.83hrs	46
SEWA	Natural	>33hrs	398
MuSe-CaR	Natural	36.87hrs	90
MSP-Conversation	Natural	15.15hrs	197

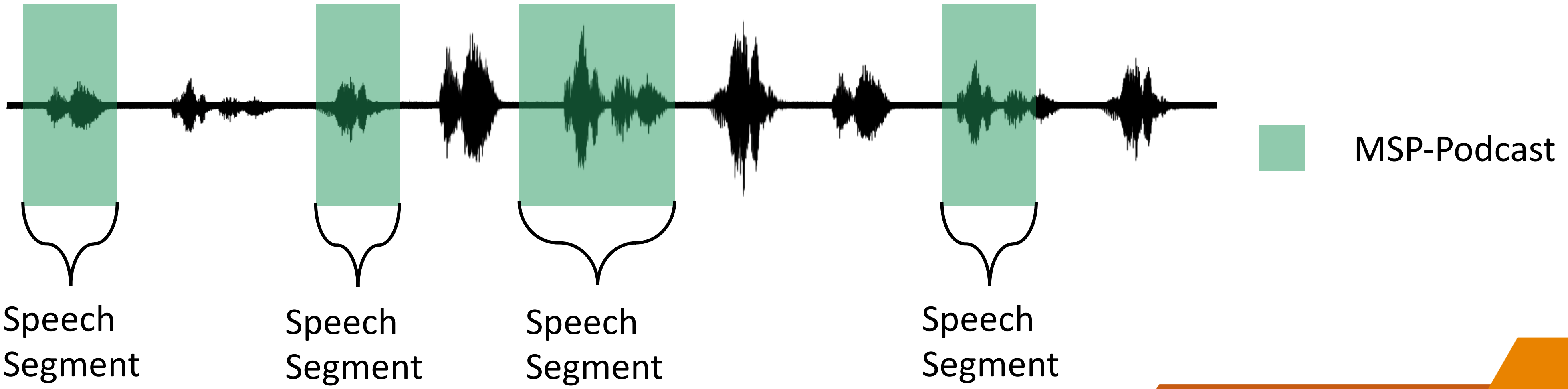
The MSP-Podcast Corpus

- **The collection of the proposed corpus is part of the efforts to collect the MSP-Podcast Corpus**
- **Speech segments from audio-sharing websites**
  - Under Creative Common licenses (CC-BY and CC-0)
  - Segments not necessarily consecutive
  - One label per segment
  - Annotated in random order



The MSP-Podcast Corpus

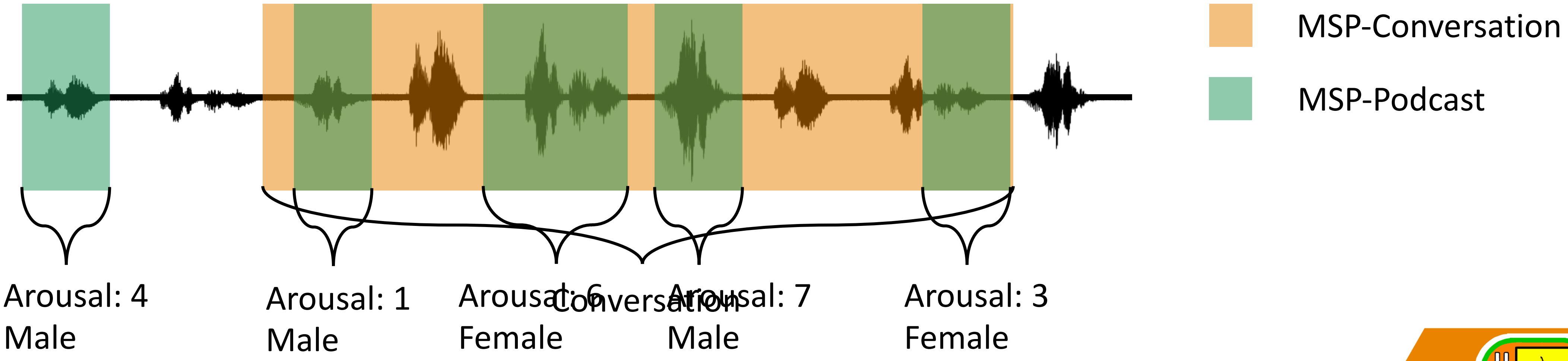
- **Limitations**
  - Cannot study effect of contextual information on emotion
  - Non-consecutive segments
  - Focus on one speaker
  - Lack of context for annotators



MSP-Conversation Corpus

Selection of Conversations

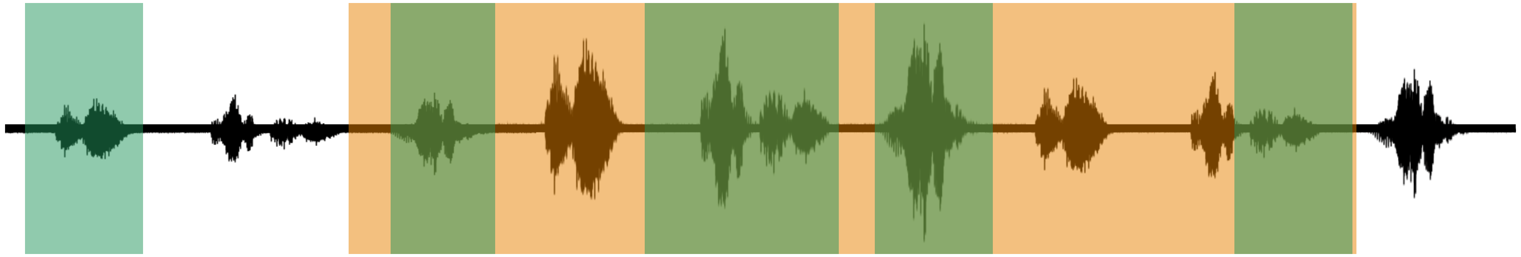
- 10 to 20-minute segments from a larger podcast
  - Natural emotional content
  - Broad range of emotions
  - Multiple speakers in spontaneous interactions
  - Balanced gender and emotional content



MSP-Conversation Corpus

- **Overlap with the MSP-Podcast Corpus**
  - Includes context in data and annotations
  - 1,567 speech segments
  - Compare sentence-level annotations with time-continuous annotations

- **Current version of the corpus**
  - 74 conversations
  - 15 hours and 9 minutes
    - 50.6% female
    - 49.4% male

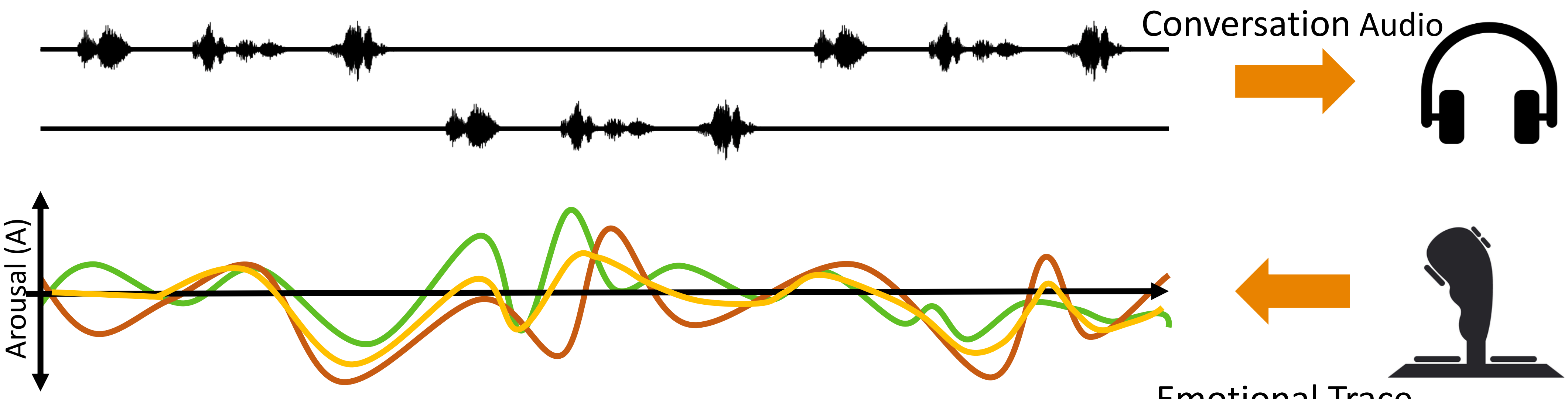


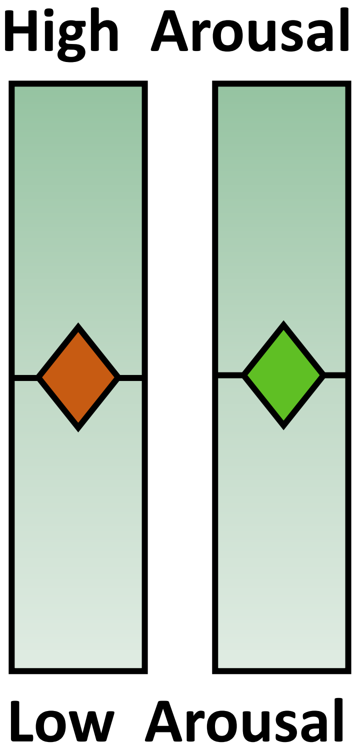
🟧 MSP-Conversation

🟩 MSP-Podcast

Annotation Protocol

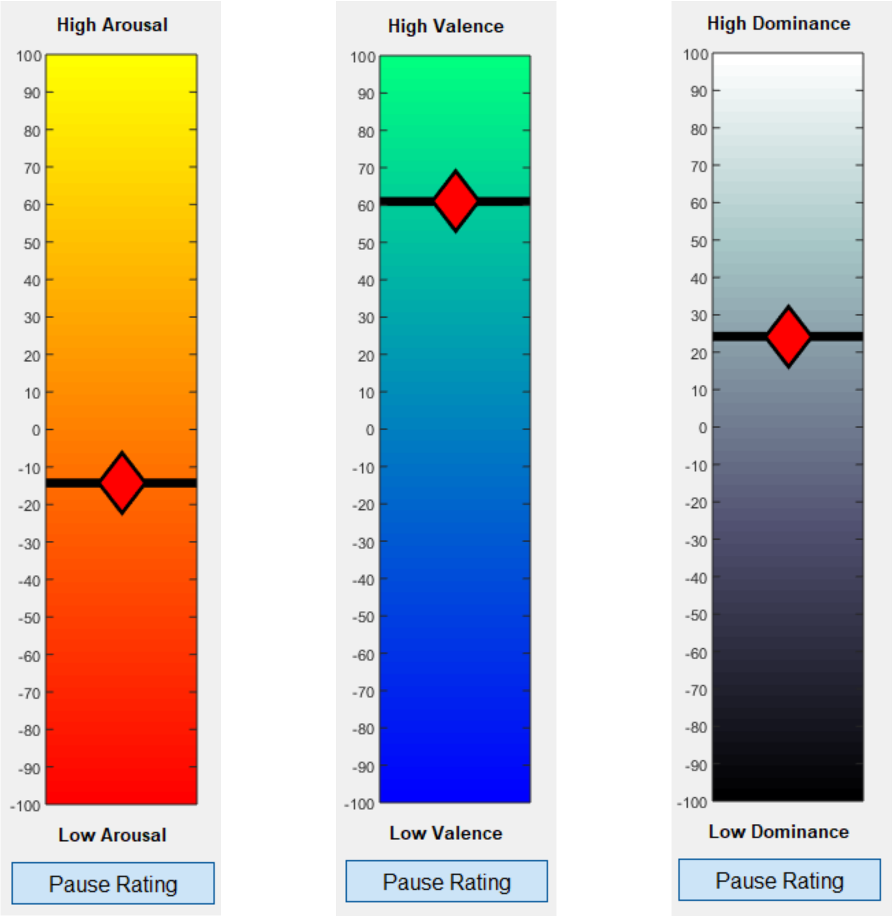
- **Conversations segmented into 3 to 7-minute segments**





- **CARMA**
  - Sampled at 60fps, but not always uniform
  - Annotated at random but in a cycle of 10 recordings with the same attribute annotations
  - 1-hour long annotation sessions with at least a 30 min break between them

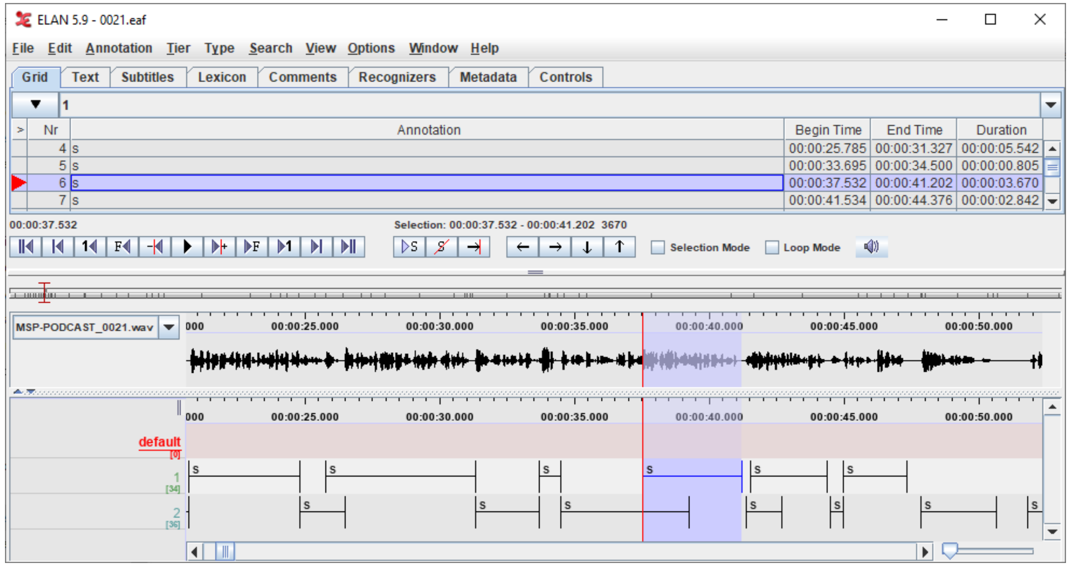
MSP-Conversation Podcast



- **We currently have 11 annotators**
- **Training of annotators**
  - Annotated 9 dialogues from the SEMAINE dataset
- **Average of annotations per conversation:**
  - 6.48 for arousal
  - 6.06 for valence
  - 5.80 for dominance
- **At least 5 annotations per conversation and attribute**

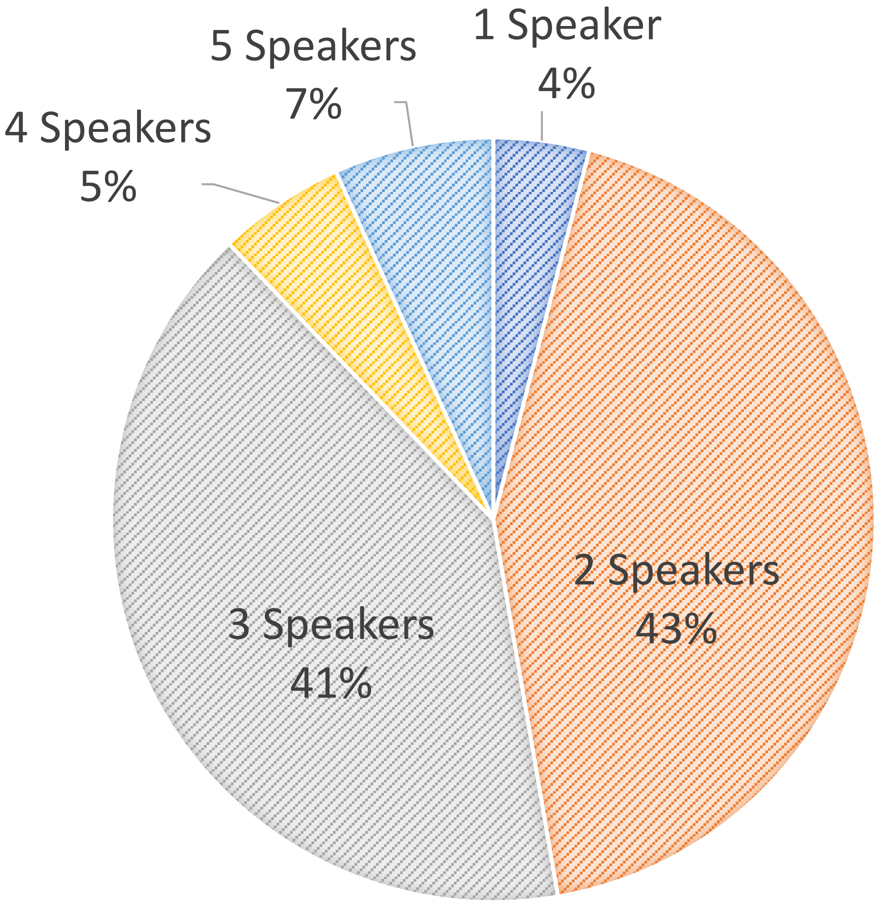
Speaker Diarization

- **Manual diarization of individual speaker activity using ELAN**
  - 197 speakers (87 female, 110 male)



ELAN 5.9 - 0021.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls
1
00:00:37.532
Selection: 00:00:37.532 - 00:00:41.202 3670
Selection Mode
Loop Mode
MSP-PODCAST_0021.wav
default
1
2

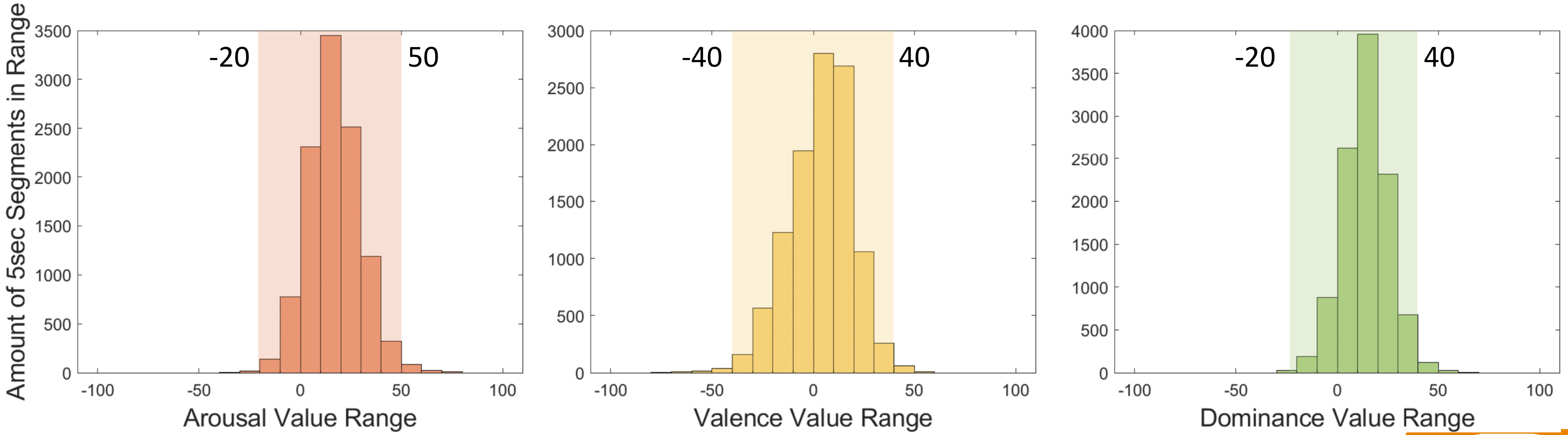
Nr	Annotation	Begin Time	End Time	Duration
4	s	00:00:25.785	00:00:31.327	00:00:05.542
5	s	00:00:33.695	00:00:34.500	00:00:00.805
6	s	00:00:37.532	00:00:41.202	00:00:03.670
7	s	00:00:41.534	00:00:44.376	00:00:02.842



SPEAKERS PER CONVERSATION

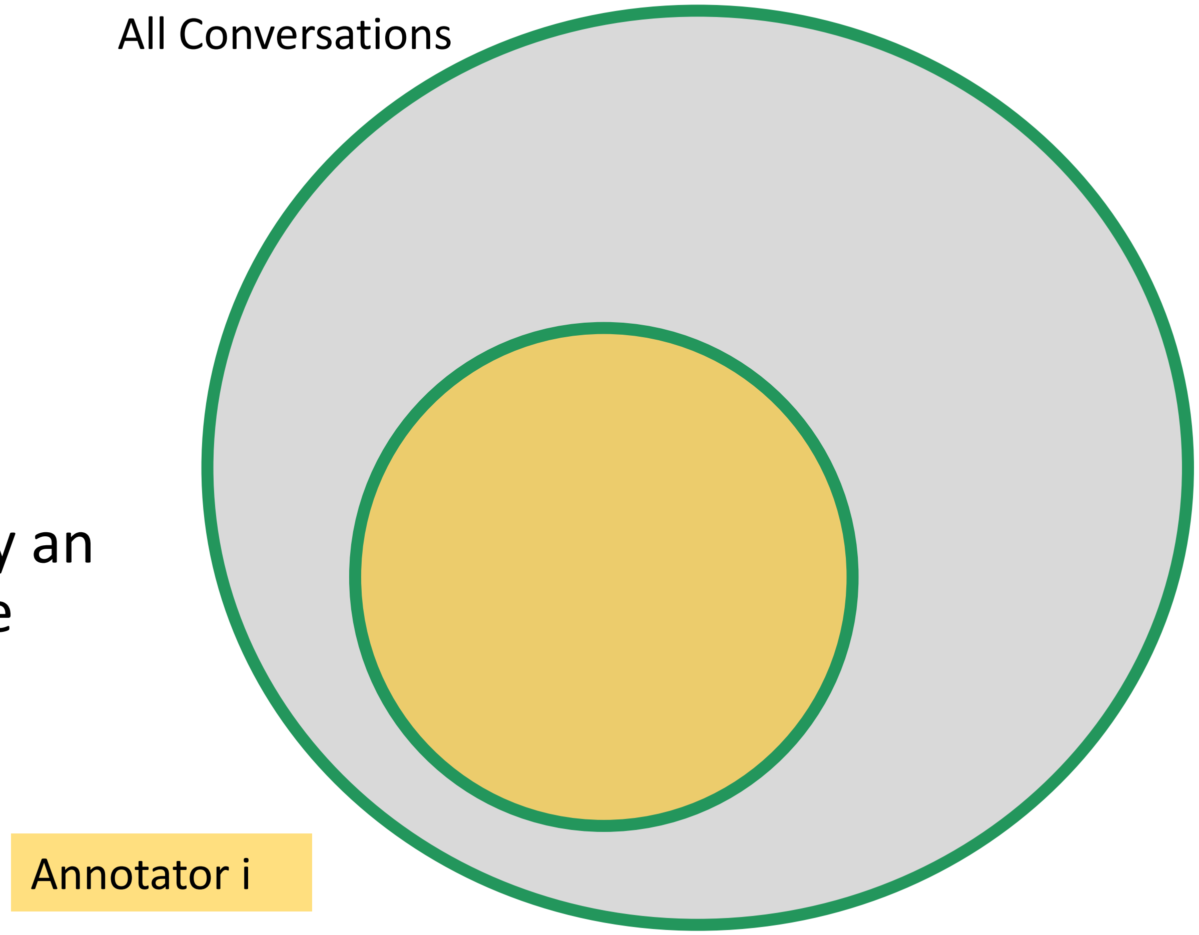
Emotional Diversity

- **Balance of emotions**
  - Valance is balanced
  - Arousal and dominance are biased towards positive values
    - Select new conversations that contain calmer and weaker behaviors



Inter-Evaluator Agreement

- **Cronbach's Alpha**
  - Measure of consistency
  - Average alpha for all conversations and each attributes
  - Average alpha for conversations annotated by an evaluator for each attribute
    - Including and excluding an evaluator

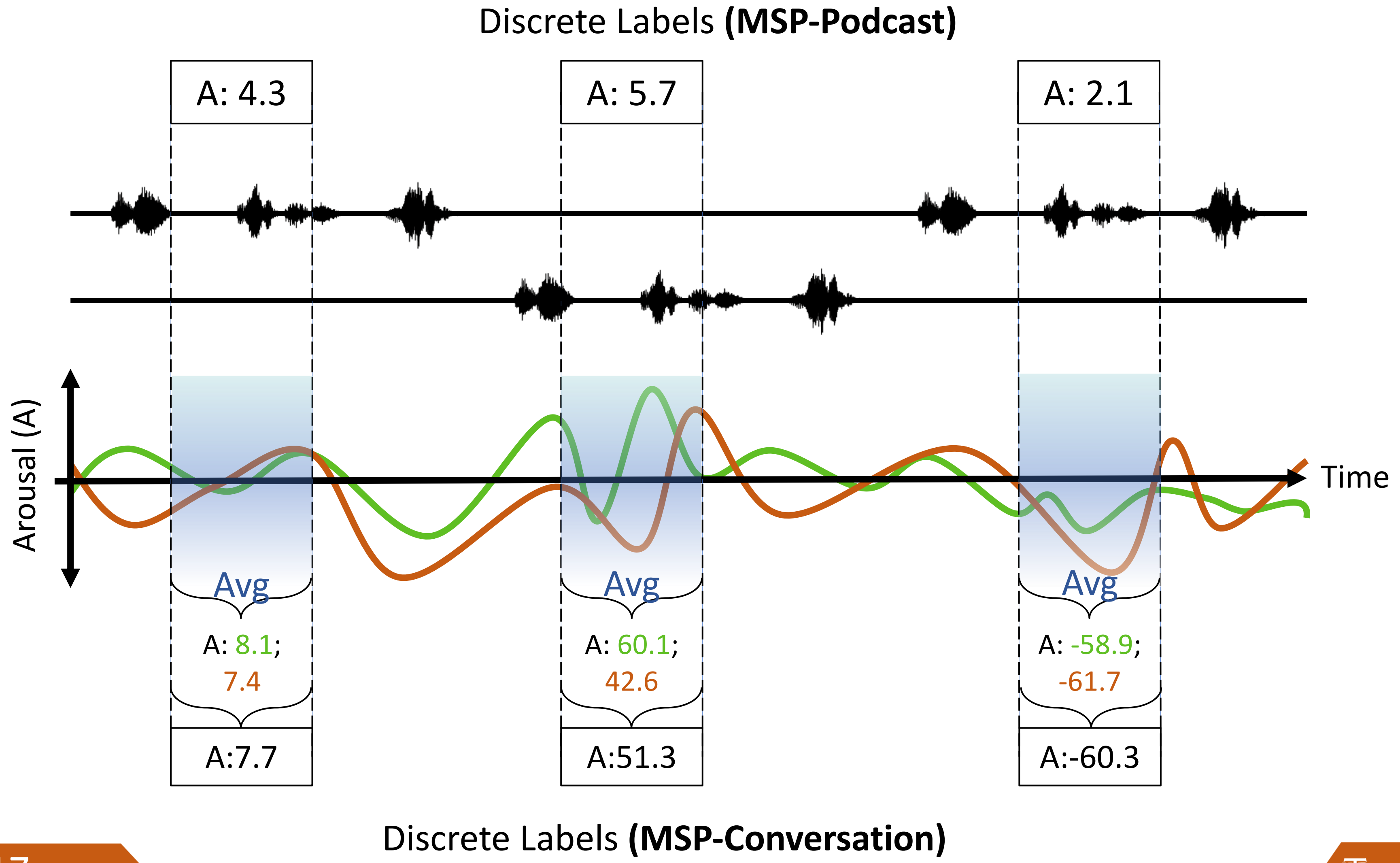


Inter-Evaluator Agreement

	Arousal		Valence		Dominance	
All	0.50		0.54		0.41	
Annotator	Included	Excluded	Included	Excluded	Included	Excluded
1	0.50	0.51	0.54	0.53	0.41	0.43
2	0.50	0.46	0.50	0.52	0.39	0.35
3	0.51	0.51	0.47	0.51	0.40	0.37
4	0.58 ←	0.53	0.63 ←	0.57	0.58 ←	0.42
5	0.50 ←	0.45	0.64 ←	0.44	0.41 ←	0.32
6	0.50 ←	0.46	0.54 ←	0.49	0.41 ←	0.37
7	0.50 ←	0.43	0.56 ←	0.51	0.44 ←	0.40
8	0.50 ←	0.41	0.56 ←	0.45	0.44 ←	0.34
9	0.56 ←	0.52	0.57 ←	0.50	0.62 ←	0.54
10	0.56 ←	0.50	0.58 ←	0.52	0.48 ←	0.40
11	0.58 ←	0.50	0.62 ←	0.59	0.54 ←	0.46

- **Agreement above α = 0.4**
- **Annotators 4 to 11 increase agreement**
  - We could weigh them more when combining traces
  - Here, we do not exclude other annotators

Time-Continuous vs Sentence-Level



- **Reaction Lag**
  - 2.8 sec
  - 3.0 sec
  - 3.6 sec
  - 4.08 sec
  - 5.44 sec
  - 5.6 sec

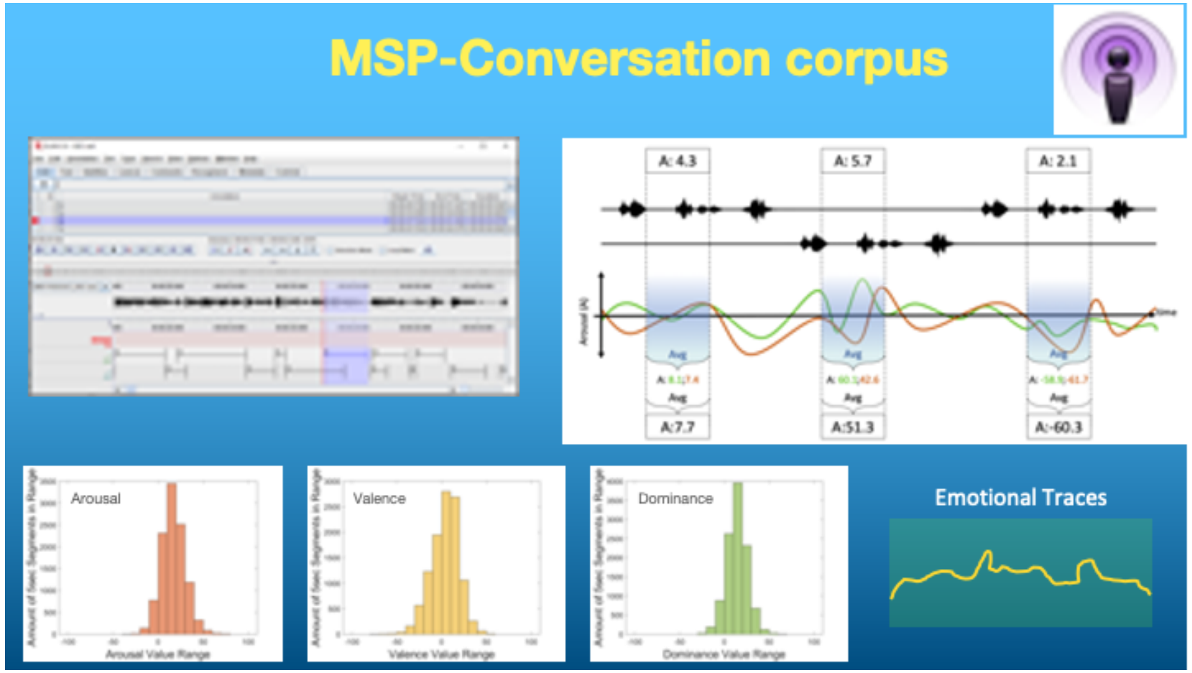
Time-Continuous vs Sentence-Level

Lag (s)	Arousal	Valence	Dominance
0.00	0.312	0.280	0.222
2.80	**0.373**	0.378	**0.273**
3.00	0.368	0.378	0.271
3.60	0.348	0.403	0.260
4.08	0.324	**0.403**	0.244
5.44	0.266	0.399	0.200
5.60	0.259	0.398	0.196

- **Pearson Correlation Coefficient**
  - Between MSP-Podcast and derived MSP-Conversation labels
  - Average coefficient for all conversations and attributes
  - Highest correlation is 0.403
    - Context makes a significant difference for evaluating emotions

Conclusion

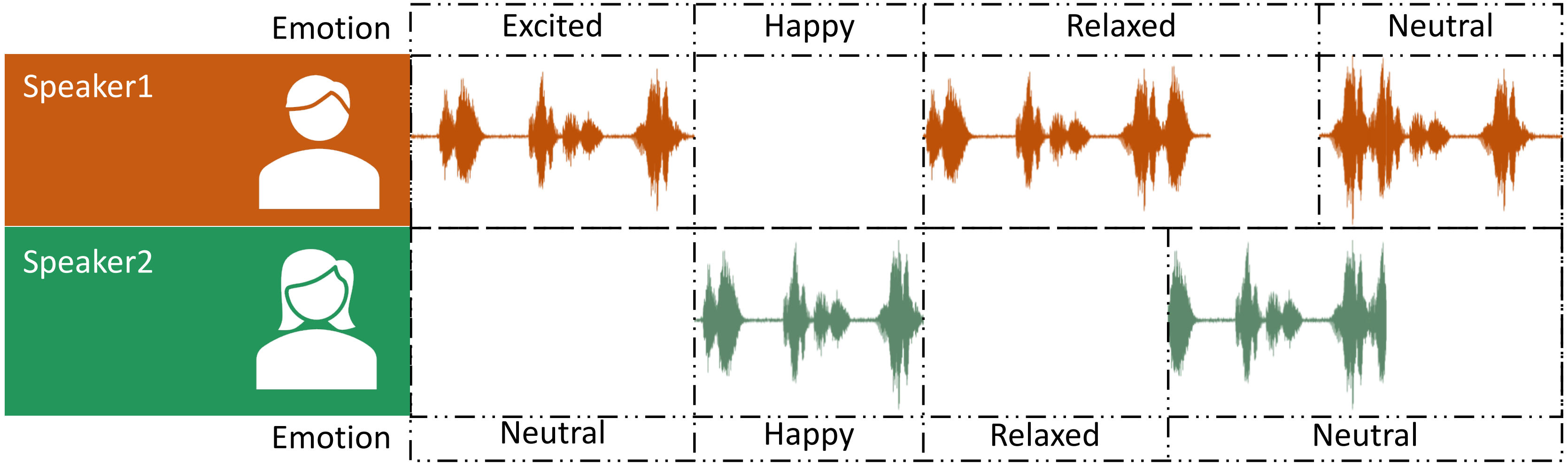
- **MSP-Conversation Corpus**
  - Time-continuous annotations
  - Naturalistic speech of multiple-party interactions
  - Scalable collection of data
  - Broad range of emotions
- **Current version of the corpus**
  - 74 conversations
  - 15 hours and 9 minutes
  - 197 speakers
  - At least 5 annotations per conversation



Future Work

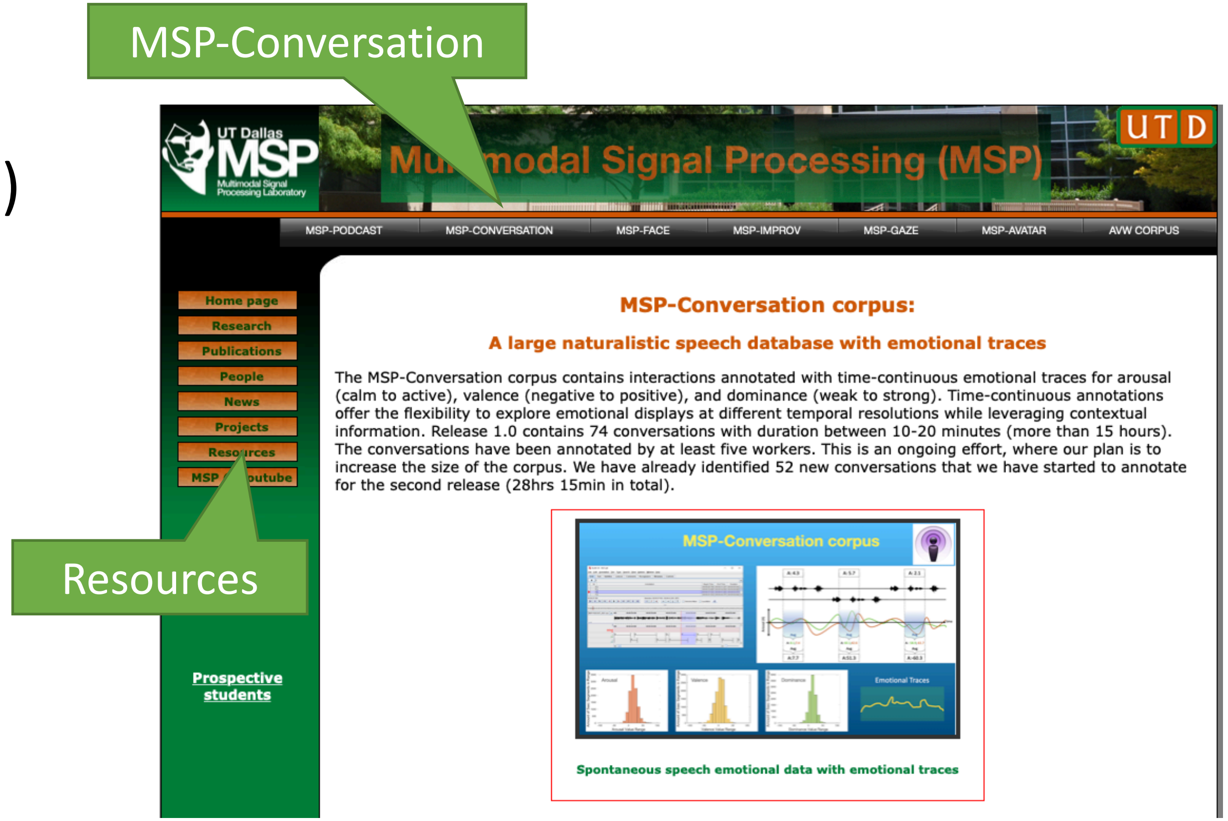
- **Ongoing effort**
  - 94 new conversations
  - 38 hours 26 minutes in total
  - Goal: 50 hours
    - At least 6 annotations per conversation

- **Future Work**
  - Analyze impact of contextual information on emotion
  - Leverage inter-dependencies between speakers in SER systems



Release of the MSP-Conversation corpus

- **Academic license**
  - Federal Demonstration Partnership (FDP) Data Transfer and Use Agreement
  - Free access to the corpus
- **Commercial license**
  - We are in the process of drafting a commercial license through UT Dallas



MSP-Conversation corpus:

A large naturalistic speech database with emotional traces

The MSP-Conversation corpus contains interactions annotated with time-continuous emotional traces for arousal (calm to active), valence (negative to positive), and dominance (weak to strong). Time-continuous annotations offer the flexibility to explore emotional displays at different temporal resolutions while leveraging contextual information. Release 1.0 contains 74 conversations with duration between 10-20 minutes (more than 15 hours). The conversations have been annotated by at least five workers. This is an ongoing effort, where our plan is to increase the size of the corpus. We have already identified 52 new conversations that we have started to annotate for the second release (28hrs 15min in total).

Spontaneous speech emotional data with emotional traces

https://msp.utdallas.edu

Thank You

- This work was funded by NSF CAREER Grant IIS-1453781

INTERSPEECH 2020

Questions or Contact: Luz Martinez-Lucas
luz.martinez-lucas@utdallas.edu

Our Research: msp.utdallas.edu